- *Social constructs*—What incentives are necessary to engage scientists in making data accessible and shared with the broader community? What are the appropriate business models necessary to promote connecting publications to data? How might private-sector participants be engaged in the effort? What are the social barriers to adopting and using unique researcher numbers?

- *The Pragmatic Experience*—What lessons have been learned from Brazil's experience with the Lattes platform? What opportunities are possible as a result of the establishment of the ORCID (Open Researcher and Contributor ID) project? What can be learned from data preservation, libraries and other coordinated data and publication efforts? What can be learned from domain-specific successes?

2. Data Access

***Exemplar*: Sloan Digital Sky Survey (SSDS)**
The SDSS is a map of the universe that was compiled from 1991 to 2008. It has generated 850 million web hits in 9 years by 1,000,000 distinct users (globally there are only 15,000 professional astronomers). SDSS tops the astronomy citation list and has delivered more than 100 billion rows of data. It has facilitated both remote collaborations and discoveries by amateur scientists (http://www.sdss.org/).

Access to data generated by the "data deluge" is crucial.[2] Research reproducibility is critical (Hirsh 2010; Donoho 2010; Donoho et al. 2009), as "[r]eplicability is a hallmark of science" (Börner 2010), but research is only reproducible if the underlying data are accessible (Stodden 2010) and reliable. The challenge is daunting: one workshop speaker noted that the scientific community now generates more data each year than the entire sum of data produced in all prior years combined (Seidel 2010). Much data are inaccessible because of the dramatic increase in the amount of "information which is 'off the records' of science, not available to peer reviewers, [and] in many cases not even recorded in formal lab notebooks or laboratory information management systems" (Pfeiffenberger 2010). A recent NSF/Office of Cyberinfrastructure (OCI) Grand Challenges Task Force Report identified reproducibility of computational results as an

[2] Recently held workshops and reports devoted to data access include the European Commission's High level Expert Group on Scientific Data, *Riding the Wave: How Europe Can Gain from the Rising Tide of Scientific Data*, October 2010, available at http://cordis.europa.eu/fp7/ict/e-infrastructure/docs/hlg-sdi-report.pdf; Yale Law School, *Data and Code Sharing Roundtable*, November 21, 2009. See also http://www.law.yale.edu/intellectuallife/codesharing.htm, and "A Special Report on Managing Information: Data, Data Everywhere," *The Economist*, February 25, 2010, available at http://www.economist.com/node/15557443.

imperative challenge for the computational sciences (National Science Foundation 2010), and a roundtable at Yale University published a declaration urging reproducible computational research through the sharing of data and code (Yale Law School Roundtable on Data and Code Sharing 2010).

"Information overload" is a related challenge, where the frequency or volume of available data overwhelms the ability of an individual or organization to usefully process, classify, manage, or analyze them (Elias 2010). Much is inaccessible due to technical difficulties accessing the proliferation of different types of data available to researchers, including numerical arrays and experimental results (Stodden 2010).

Participants noted that funding agencies should act due to abundant evidence that data access is valuable in advancing science (Raddick and Szalay 2010; Donoho et al. 2009). This includes remote collaborations among scientists, such as those using data available through the Sloan Digital Sky Survey (see Exemplar) (Raddick and Szalay 2010; Neylon 2010a), European Supersites for Atmospheric Aerosol Research (EUSAAR), or biodiversity research (Wood 2010). A cyber community has collaborated on infrastructure development to advance biodiversity research. The European Space Agency validates satellite Earth observation data, CERN (The European Organization for Nuclear Research) enables the grids for e-science, and the Global Biodiversity Information Facility (GBIF) and Encyclopedia of Life (EOL) facilitate data access (Raddick and Szalay 2010). Data accessibility has also helped citizen science flourish (Raddick and Szalay 2010; Hirsh 2010; Stodden 2010), and it has contributed to the concept of "collective intelligence," which "seeks to understand [the] new ways in which people collaborate and create outcomes that are integrally about large groups of participating individuals, as much as they are about the new technologies that underlie them" (Hirsh 2010). Academic research may also be enriched by greater access to the abundance of data already being collected in the public and private sectors (German Data Forum (RatSWD) 2010), provided that privacy and proprietary rights are protected.

The workshop participants discussed the social and technical challenges associated with promoting data access.

2.1. Social Challenges

Workshop participants agreed that the primary social barriers associated with data access include insufficient intellectual property rights, the difficulty of documenting data for reuse, and the problems associated with protecting confidentiality and privacy.

There was agreement that academic institutions do not completely recognize the ownership and intellectual property rights relating to data production and sharing. There are also legal constraints: current copyright and other intellectual property laws in many nations present legal barriers to fully sharing data, articles, papers, methodologies, and code.

Workshop participants also noted that any comprehensive data-access plan must resolve the tension between confidentiality and openness (Schutz 2010). As the European Union has acknowledged, “[i]nnovation is important in today’s society, but should not go at the expense of people’s fundamental right to privacy”(EU NewsBrief 2010). While the German Data Forum has recognized the enormous research potential of allowing access to official census data and other sources of public data, it has also emphasized respect for individual privacy and the need to protect individually identifiable data (German Data Forum (RatSWD) 2010). Legal, ethical, and administrative restrictions on the reuse of data sets containing personally identifiable information, such as income, health, and criminal records (Elias 2010), are intended to safeguard human subject privacy, ensure subject consent for the use of personal data, or provide stewardship. These restrictions are often applied to data sets gathered and maintained by national statistical offices and agencies.

Participants made a number of suggestions for addressing social challenges related to data access. “Persuasive” incentives, such as attribution or linking data sets to subsequent publications, are needed to encourage researchers to give the wider research community access to their data (Pfeiffenberger 2010). Moreover, it is important to “value the publication of data (and software) as potentially equivalent to articles about conclusions, methods, instrumentation, models, algorithms and whatever is considered a legitimate object of publication” (Pfeiffenberger 2010). Published data could serve as an assessment and certification of quality, much as the publication of a peer-reviewed academic article represents the vetting of an argument or concept, and allow data sets to become part of “the scientific record” (Pfeiffenberger 2010). In addition, researchers that have openly provided their data to others should be recognized through attribution in any subsequent publication that makes use of the data (Trasande and Hannay 2010).

***Exemplar*: Earth System Science Data** The Data Publishing Journal provides quality assessments for data sets that reside in permanent repositories. The journal maps peer-review criteria from text to data (http://www.earth-system-science-data.net/review/ms_evaluation_criteria.html).

One option for resolving the conflicts between reproducibility and copyright law is the development of an Open Research License (ORL) to “encourage researchers to create fully reproducible research by allowing [them] to capture more of the credit for facilitating and expanding scientific understanding, while promoting the ideal of reproducible research” (Stodden 2008).

Participants agreed that training is critical to maintaining open data systems, whether informal, formal, or through discussions and research practices. Training could take place at the graduate and post-doc levels, potentially instilling good habits at an early career stage. Colleges and universities could require graduate students and post docs to write papers using only publicly available data and publish them in open-source journals [Schutz; Trasande].[3] But without advisors setting an example by using open data, students would not see it as a laudable practice [Sauermann]. It is also important to develop a method for crediting informal training and learning to provide incentives to perpetuate the education of successive generations of researchers [Trasande].

One potential method for enhancing access to data while maintaining human subject privacy is the use of "virtual safe settings—systems through which authorised and authenticated users can gain remote access to data on individuals or organisations whilst preventing copying of data and minimising the potential for abuse of access privileges" (Elias 2010).

***Exemplar*: Permanent Access to the Records of Science in Europe (PARSE) Insight**

The EU is funding PARSE, is a 2-year project focused on the preservation of digital information in science over time and ensuring that it "is accessible, usable and understandable in the future." The goal is to create a roadmap for facilitating continuous access to scientific data (http://www.parse-insight.eu/).

2.2. Technical Issues

Workshop participants noted that increasing access to research data involves solving a host of technical issues, including data control, security, long-term data preservation, and stewardship.

Data storage should be planned from the start of any data-sharing enterprise (Wood 2010), and access control, archive security, and protection of confidential data should also be considered during the planning process (Schutz 2010). In this sense, the Permanent Access to the Records of Science in Europe Insight project (see Exemplar) serves as a potential model for scientific data infrastructure (Wood 2010). Providing useful access to the magnitude of research data that is continually being created is a primary challenge of any effort to create and harmonize a global scientific data infrastructure. One participant pointed out that scope presents the greatest barrier, asserting, "petabytes are easy, exabytes are hard" [Hirsh]. Not only must data be collected and stored, they must also be retrievable in discrete sets that can easily be reused by researchers.

Several options for addressing technical issues related to data access were discussed by participants. An important conceptual framework for creating a user-friendly scientific

[3] Last names in brackets refer to participants who made comments during the workshop. *See* Appendix C for participant biographies.

data infrastructure may include a “Knowledge Organisation System” that provides a consistent means of describing science, maintains an overview of the interrelationship between various areas of scientific knowledge, and presents an extract of the connections between past scientific knowledge and the emergence of new scientific knowledge from current and future research (Lambe 2010). Another possible model is a Reproducible Research System (RRS), which has two parts: (1) a Reproducible Research Environment (RRE) for computation work, which “provides computational tools together with the ability to automatically track the provenance of data, analyses, and results, and to package them (or pointers to persistent versions of them) for redistribution,” and (2) a Reproducible Research Publisher (RRP), such as standard word-processing software or other documentation-preparation system, that links to the RRE. This facilitates readers’ abilities to reproduce the analysis, as well as “extend it with the document itself by changing parameters, data, filters, and so on” (Pfeiffenberger 2010).

“Data...needs to be accessible by anyone, from anywhere, at anytime” (Viegas 2010). This requires the creation of taxonomies of scientific data to enable the cataloguing, tagging, and parsing of data sets for automated recall. Currently, a number of scientific data infrastructure systems use different and incompatible data identifiers, inhibiting data sharing and reuse (Fenner 2010). Another obstacle to the creation of a useful scientific data sharing infrastructure is the issue of interoperability—ensuring that researchers can easily reuse data sets originating in any country. To overcome this issue, international data standards and taxonomies must be explored by the scientific research community. This will likely first occur in individual disciplines, then in interdisciplinary conversations.

Standardized identification schemes, such as Altman and King’s Universal Numerical Fingerprint implemented at the Dataverse Network at Harvard University (Altman and King 2007); metadata standards like MIAME for microarray gene expression (Trasande and Hannay 2010; Fenner 2010); and Digital Object Identifiers (DOIs) for any physical or digital manifestation, including text, audio, images, and software (Fenner 2010), can aid in categorizing and managing data and data sources and increase interoperability and ease of use. However, for these schemes to be successful, scientists must be encouraged and given incentives to routinely use the standards to annotate their data, a process that will be aided by the development of better and easier to use software tools (Trasande and Hannay 2010).

Not only do all students need to be trained to utilize open data, but individuals need to be formally educated as “data scientists” [Wood]. A new cohort of computational scientists who can manage the integration of data sets from disparate sources is essential [Aragão; Santos].

2.3. Role of Funding Agencies

Workshop participants suggested a number of ways in which funding agencies worldwide can significantly improve data access. For example, they can provide incentives for data sharing by publishing openness rankings [Evans], or using curriculum grants or similar measures to encourage informal and formal training of students on the importance of open data [Seidel]. Agencies can promote interoperability through international initiatives to develop persistent digital research data infrastructure.

Current efforts include, the U.S. National Science and Technology Council Interagency Working Group on Digital Data's recommendation that all U.S. federal agencies "promote a data management planning process for projects that generate scientific data for preservation" (Office of Science and Technology Policy 2009). NSF has complied with this call to action and changed the implementation of its long-standing data policy[4] by requiring that beginning in January 2011, all proposals include a "Data Management Plan" that describes:

- The types of data, samples, physical collections, software, curriculum materials, and other materials to be produced in the course of the project;
- The standards to be used for data and metadata format and content (where existing standards are absent or deemed inadequate, this should be documented along with any proposed solutions or remedies);
- Policies for access and sharing including provisions for appropriate protection of privacy, confidentiality, security, intellectual property, or other rights or requirements;
- Policies and provisions for reuse, redistribution, and the production of derivatives; and
- Plans for archiving (National Science Foundation 2011).

Other funding agencies have also taken an active role in ensuring data access. For example, the Alliance of German Science Organisations published the June 2008 "Digital Information Initiative" designed "to equip scientists and academics with the information and infrastructures best suited to facilitate their scientific work" (Lauer 2010).[5] The United Kingdom's Economic and Social Research Council website (http://www.esrc.ac.uk/about-esrc/information/data-policy.aspx) also notes:

[4] "Investigators are expected to share with other researchers, at no more than incremental cost and within a reasonable time, the primary data...created or gathered in the course of work under NSF grants" (National Science Foundation 2010a).

[5] *See* also Deutsche Forschungs Gemeinschaft, http://www.dfg.de/en/research_funding/programmes/infrastructure/lis/digital_information/alliance_initiative/index.html.

> [T]hose ESRC grant applicants who plan to generate data are responsible for preparing and submitting data management and sharing plans for their research projects as an integral part of the application. It is then a responsibility of the award holder to incorporate data management and sharing as an indivisible part of the research project to increase the potential for data to be shared. We require that the data must be made available for preparation for [reuse] and/or archiving with the ESRC data service providers within three months of the end of the award otherwise we will withhold the final payment.

3. Knowledge Access

Sharing knowledge about scientific discoveries is a foundation of modern science, but workshop participants noted that funding agencies need to understand that knowledge, and therefore knowledge sharing, should be broadly defined to encompass both data and code. They also noted that knowledge sharing takes many forms and should be encouraged, including traditional academic journal publishing as well as other mechanisms such as discussion forums, recommendations, wikis, file-sharing sites, blogs, and microblogs (Trasande and Hannay 2010). However, substantial barriers to knowledge access persist despite mandates to promote sharing. For example, in spite of the embrace of Open Access publishing, the voluntary adoption rate by scientists has been low (around 15%–20%). Mandates have increased these numbers to around 70% for NIH-funded research and in institutions, such as Southampton or CERN, that have adopted these policies. Nevertheless, this means that even with mandatory participation, some 30% of research is not openly available (Fenner 2010).

3.1. Social Issues

Workshop participants agreed that attribution for new forms of scientific activity was critical to promoting knowledge access. Researchers will provide access to their work if they are given credit for their labor. Attribution for scholarly work requires the ability to uniquely identify both specific contributors to research and specific scientific contributions (Fenner 2010). Participants felt strongly that an author-identification system that transcends institutional, disciplinary, and national boundaries would help create a "clear and unambiguous scholarly record" of research activities associated with an individual and help provide unambiguous attribution for researcher contributions, whether they appear as publications, patents, or data sets (Office of Science and Technology Policy 2009; National Science Foundation 2011). An author-identification system would also allow for "microattribution" for research contributions not associated with a peer-reviewed journal publication (Credit Where Credit Is Due 2009). In the current system, a significant portion of scientific work remains unrecognized because there are no formal methods for providing attribution for this labor